Systematic sampling locations for comparing a median with a fixed threshold (nonparametric - MARSSIM)

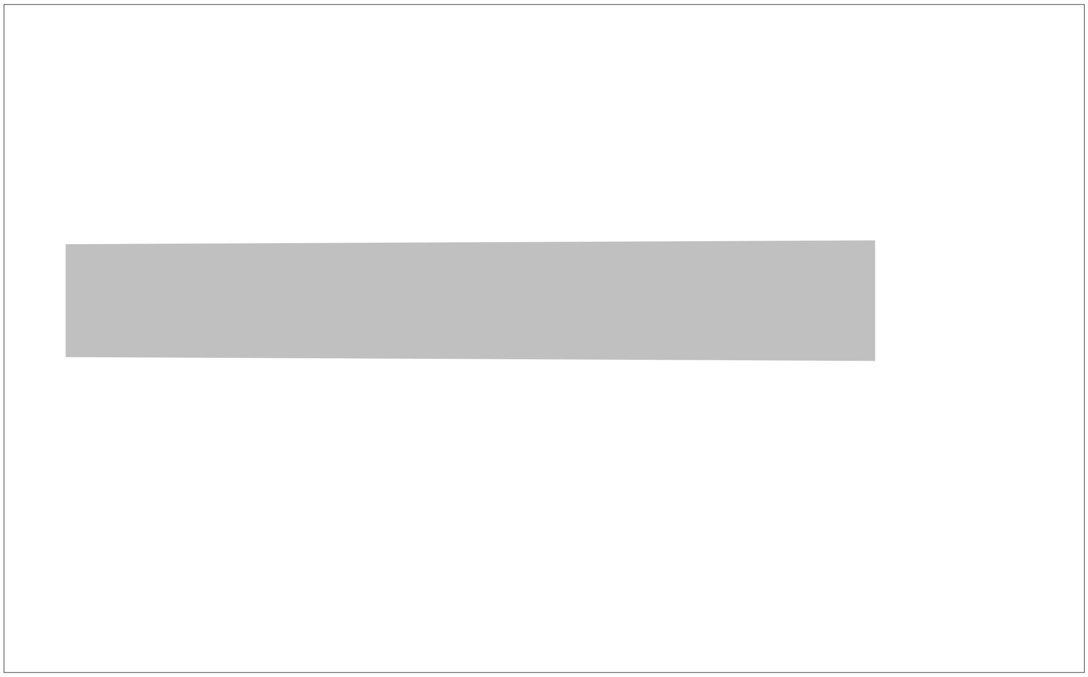
Summary

This report summarizes the sampling design used, associated statistical assumptions, as well as general guidelines for conducting post-sampling data analysis. Sampling plan components presented here include how many sampling locations to choose and where within the sampling area to collect those samples. The type of medium to sample (i.e., soil, groundwater, etc.) and how to analyze the samples (in-situ, fixed laboratory, etc.) are addressed in other sections of the sampling plan.

The following table summarizes the sampling design developed. A figure that shows sampling locations in the field and a table that lists sampling location coordinates are also provided below.

SUMMARY OF SAMPLING DESIGN	
Primary Objective of Design	Compare a site mean or median to a fixed threshold
Type of Sampling Design	Nonparametric
Sample Placement (Location) in the Field	Systematic with a random start location
Working (Null) Hypothesis	The median(mean) value at the site exceeds the threshold
Formula for calculating number of sampling locations	Sign Test - MARSSIM version
Calculated total number of samples	9
Number of samples on map [a]	10
Number of selected sample areas [b]	1
Specified sampling area [c]	0.00 ft^2
Size of grid / Area of grid cell [d]	6.15381e-005 x 0.000184614 feet / 1.13608e-008 ft^2
Grid pattern	Rectangular

[a] This number may differ from the calculated number because of 1) grid edge effects, 2) adding judgment samples, or 3) selecting or unselecting sample areas.
[b] The number of selected sample areas is the number of colored areas on the map of the site. These sample areas contain the locations where samples are collected.
[c] The sampling area is the total surface area of the selected colored sample areas on the map of the site.
[d] Size of grid / Area of grid cell gives the linear and square dimensions of the grid used to systematically place samples.
[e] Including measurement analyses and fixed overhead costs. See the Cost of Sampling section for an explanation of the costs presented here.



Area: Area 1					
X Coord	**Y Coord**	**Label**	**Value**	**Type**	**Historical**
119.639817	46.559671	FS-1-2		Systematic	
119.640002	46.559671	FS-1-4		Systematic	
119.640186	46.559671	FS-1-6		Systematic	
119.640371	46.559671	FS-1-8		Systematic	
119.640556	46.559671	FS-1-10		Systematic	
119.639817	46.559732	FS-1-1		Systematic	
119.640002	46.559732	FS-1-3		Systematic	
119.640186	46.559732	FS-1-5		Systematic	
119.640371	46.559732	FS-1-7		Systematic	
119.640556	46.559732	FS-1-9		Systematic	

Primary Sampling Objective
The primary purpose of sampling at this site is to compare a site median or mean value with a fixed threshold. The working hypothesis (or 'null' hypothesis) is that the median(mean) value at the site is equal to or exceeds the threshold. The alternative hypothesis is that the median(mean) value is less than the threshold. VSP calculates the number of samples required to reject the null hypothesis in favor of the alternative one, given a selected sampling approach and inputs to the associated equation.

Selected Sampling Approach
A nonparametric systematic sampling approach with a random start was used to determine the number of samples and to specify sampling locations. A nonparametric formula was chosen because the conceptual model and historical information (e.g., historical data from this site or a very similar site) indicate that typical parametric assumptions may not be true.

Both parametric and non-parametric equations rely on assumptions about the population. Typically, however, non-parametric equations require fewer assumptions and allow for more uncertainty about the statistical distribution of values at the site. The trade-off is that if the parametric assumptions are valid, the required number of samples is usually less than if a non-parametric equation was used.

Locating the sample points over a systematic grid with a random start ensures spatial coverage of the site. Statistical analyses of systematically collected data are valid if a random start to the grid is used. One disadvantage of systematically collected samples is that spatial variability or patterns may not be discovered if the grid spacing is large relative to the spatial patterns.

Number of Total Samples: Calculation Equation and Inputs
The equation used to calculate the number of samples is based on a Sign test (see PNNL 13450 for discussion). For this site, the null hypothesis is rejected in favor of the alternative one if the median(mean) is sufficiently smaller than the threshold. The number of samples to collect is calculated so that if the inputs to the equation are true, the calculated number of samples will cause the null hypothesis to be rejected.

The formula used to calculate the number of samples is:

$$n = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{4(SignP - 0.5)^2}$$

where

$$SignP = \Phi\left(\frac{\Delta}{s_{total}}\right)$$

$\Phi(z)$ is the cumulative standard normal distribution on $(-\infty,z)$ (see PNNL-13450 for details),
n is the number of samples,
S_{total} is the estimated standard deviation of the measured values including analytical error,
Δ is the width of the gray region,
α is the acceptable probability of incorrectly concluding the site median(mean) is less than the threshold,
β is the acceptable probability of incorrectly concluding the site median(mean) exceeds the threshold,
$Z_{1-\alpha}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\alpha}$ is $1-\alpha$,
$Z_{1-\beta}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\beta}$ is $1-\beta$.

Note: MARSSIM suggests that the number of samples should be increased by at least 20% to account for missing or unusable data and uncertainty in the calculated value of n. VSP allows a user-supplied percent overage as discussed in MARSSIM (EPA 2000, p. 5-33).

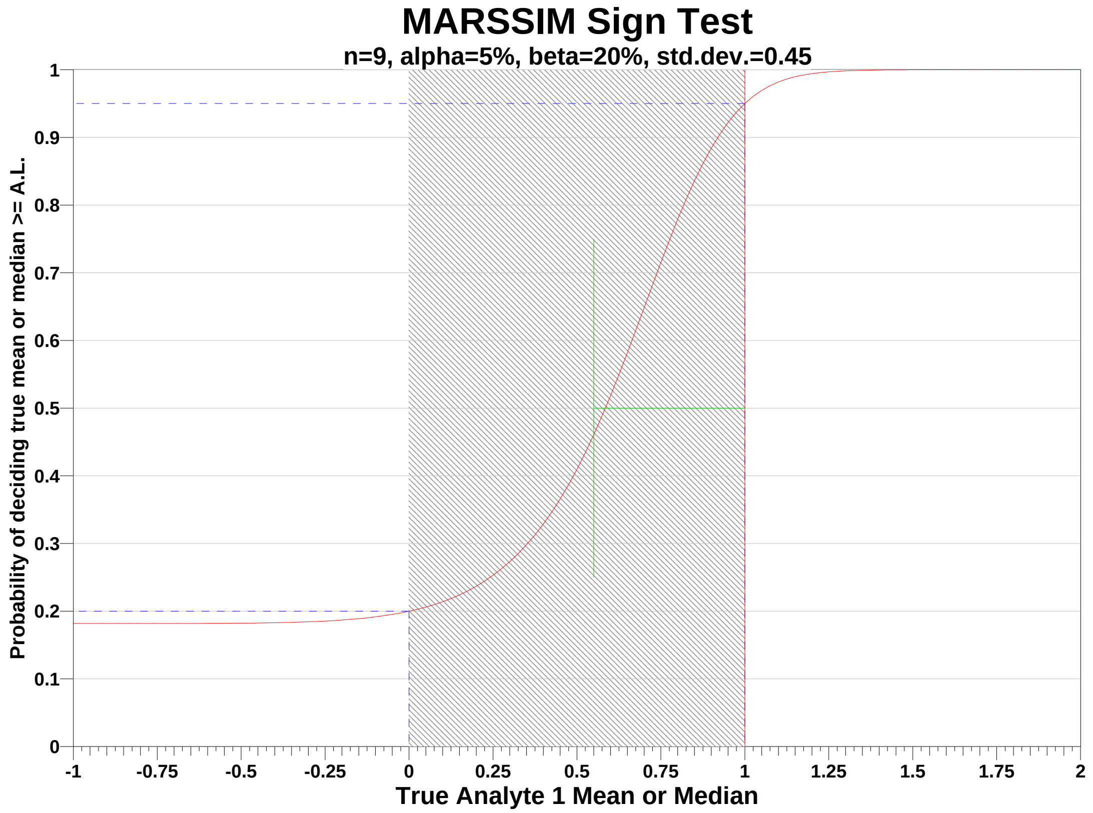
The values of these inputs that result in the calculated number of sampling locations are:

Analyte	n[a]	Parameter					
		S	Δ	α	β	$Z_{1-\alpha}$ [b]	$Z_{1-\beta}$ [c]
Analyte 1	9	0.45	1	0.05	0.2	1.64485	0.841621

[a] The final number of samples has been increased by the MARSSIM Overage of 20%.
[b] This value is automatically calculated by VSP based upon the user defined value of α.
[c] This value is automatically calculated by VSP based upon the user defined value of β.

The following figure is a performance goal diagram, described in EPA's QA/G-4 guidance (EPA, 2000). It shows the probability of concluding the sample area is dirty on the vertical axis versus a range of possible true median(mean) values for the site on the horizontal axis. This graph contains all of the inputs to the number of samples equation and pictorially represents the calculation.

The red vertical line is shown at the threshold (action limit) on the horizontal axis. The width of the gray shaded area is equal to Δ; the upper horizontal dashed blue line is positioned at $1-\alpha$ on the vertical axis; the lower horizontal dashed blue line is positioned at β on the vertical axis. The vertical green line is positioned at one standard deviation below the threshold. The shape of the red curve corresponds to the estimates of variability. The calculated number of samples results in the curve that passes through the lower bound of Δ at β and the upper bound of Δ at $1-\alpha$. If any of the inputs change, the number of samples that result in the correct curve changes.



Statistical Assumptions

The assumptions associated with the formulas for computing the number of samples are:

1. the computed sign test statistic is normally distributed,
2. the variance estimate, S^2, is reasonable and representative of the population being sampled,
3. the population values are not spatially or temporally correlated, and
4. the sampling locations will be selected probabilistically.

The first three assumptions will be assessed in a post data collection analysis. The last assumption is valid because the gridded sample locations were selected based on a random start.

Sensitivity Analysis

The sensitivity of the calculation of number of samples was explored by varying the standard deviation, lower bound of gray region (% of action level), beta (%), probability of mistakenly concluding that μ > action level and alpha (%), probability of mistakenly concluding that μ < action level. The following table shows the results of this analysis.

Number of Samples							
AL=1		**α=5**		**α=10**		**α=15**	
		s=0.9	**s=0.45**	**s=0.9**	**s=0.45**	**s=0.9**	**s=0.45**
LBGR=90	**β=15**	1103	280	825	209	659	167
	β=20	948	240	692	176	542	138
	β=25	826	209	587	149	449	114
LBGR=80	**β=15**	280	75	209	56	167	45
	β=20	240	64	176	47	138	36
	β=25	209	56	149	40	114	30
LBGR=70	**β=15**	128	36	95	27	77	22
	β=20	110	32	81	23	63	18
	β=25	95	27	69	20	52	15

s = Standard Deviation
LBGR = Lower Bound of Gray Region (% of Action Level)
β = Beta (%), Probability of mistakenly concluding that μ > action level
α = Alpha (%), Probability of mistakenly concluding that μ < action level
AL = Action Level (Threshold)

Recommended Data Analysis Activities

Post data collection activities generally follow those outlined in EPA's Guidance for Data Quality Assessment (EPA, 2000). The data analysts will become familiar with the context of the problem and goals for data collection and assessment. The data will be verified and validated before being subjected to statistical or other analyses. Graphical and analytical tools will be used to verify to the extent possible the assumptions of any statistical analyses that are performed as well as to achieve a general understanding of the data. The data will be assessed to determine whether they are adequate in both quality and quantity to support the primary objective of sampling.

Because the primary objective for sampling for this site is to compare the site median(mean) value with a threshold value, the data will be assessed in this context. Assuming the data are adequate, at least one statistical test will be done to perform a comparison between the data and the threshold of interest. Results of the exploratory and quantitative assessments of the data will be reported, along with conclusions that may be supported by them.

This report was automatically produced* by Visual Sample Plan (VSP) software version 6.5.
Software and documentation available at http://vsp.pnnl.gov

* - The report contents may have been modified or reformatted by end-user of software.